(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
2 October 2003 (02.10.2003)

PCT

(10) International Publication Number
WO 03/081416 A2

(51) International Patent Classification[7]: G06F 3/06, 1/32

(21) International Application Number: PCT/US03/08864

(22) International Filing Date: 21 March 2003 (21.03.2003)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/366,202 21 March 2002 (21.03.2002) US

(71) Applicant: TEMPEST MICROSYSTEMS [US/US]; Suite 110, 2223 Avenida de la Playa, San Deigo, CA 92037 (US).

(72) Inventors: FISK, Ian; 5730 Ferber Street, San Diego, CA 92122 (US). MOJAVER, Michael; 5194 Manor Ridge Lane, San Diego, CA 92130 (US).

(74) Agents: MOLLAAGHABABA, Reza et al.; Nutter, McClennen & Fish LLP, World Trade Center West, 155 Seaport Boulevard, Boston, MA 02210-2604 (US).

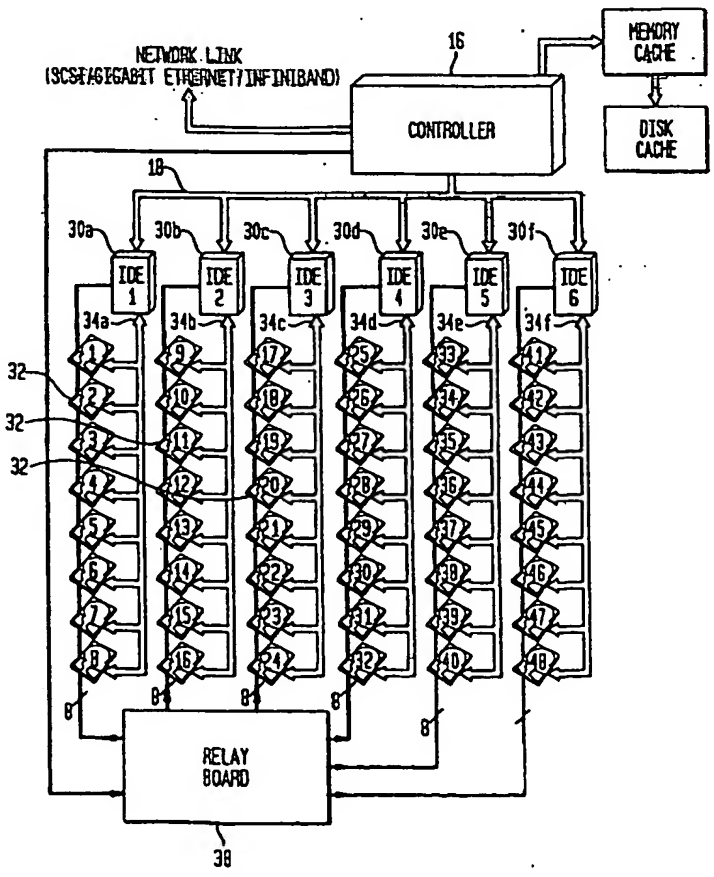
(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NI, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

**Published:**
— *without international search report and to be republished upon receipt of that report*

*[Continued on next page]*

(54) Title: A LOWER POWER DISK ARRAY AS A REPLACEMENT FOR ROBOTIC TAPE STORAGE




(57) Abstract: The present invention provides methods and systems for storage of data. In one aspect, the invention provides data storage system that includes a plurality of storage devices, such as, disks, for storing data, and a controller that implements a policy for managing distribution of power to the storage devices, which are normally in a power-off mode. In particular, the controller can effect transition of a storage device from a power-off-mode to a power-on mode upon receipt of a request for reading data from or writing data to that storage device. The controller further effects transition of a storage device from a power-on mode to a power-off mode if no read/write request is pending for that storage device and a selected time period, *e.g.*, a few minutes, has elapsed since the last read/write request for that storage device.

WO 03/081416 A2

# A LOWER POWER DISK ARRAY AS A REPLACEMENT FOR ROBOTIC TAPE STORAGE

## Related Applications

The present application claims priority to a provisional application entitled "A lower power disk array as a replacement for robotic tape storage" filed on March 21, 2002 and having Application Serial No. 60/366202. This provisional application is herein incorporated by reference.

## Background

The present invention relates to methods and systems for storing data, and more particularly, to cost-effective methods and systems for storage and retrieval of a large amount of data, e.g., in a range of tens to hundreds of Terabytes.

The volume of data generated by business processes in variety of organizations is increasing exponentially with time. Most industrial and business processes are far more efficient in generating digital data than in utilizing it. As a result, the demand for long-term data storage and back-up is growing rapidly. Currently, large scale data warehousing is typically implemented by employing tape media, which suffer from long access latency, namely, the time required for loading the tape and other associated access times. In addition, robotic tape systems are bulky and expensive to maintain.

Since the latency period for access to database items located in a tape archive is typically on the scale of tens to hundreds of seconds, a system overload frequently arises when a database search requires access to data located on many or all of the tapes in a library. Improving robotic tape storage access presents a challenging problem. Even with multiple arms and tape drives, access within each tape remains serial with few opportunities for speeding up access to data. Soft-ware approaches that streamline tape access by clustering and de-clustering multiple accesses are known. These approaches can improve performance of Petabyte tape libraries that include several hundred Terabytes of disk cache. These approaches, however, can not eliminate the fundamental limitations arising from tape access latency.

The controller can be implemented as a central device to manage power distribution to all storage devices in a manner described above. Alternatively, a plurality of controllers, each managing power distribution to each individual storage device or a group of storage devices, can be employed. Hence, the term "controller," as used herein, is intended to refer to a single central control device or a plurality of devices that collectively implement a policy for distributing power to a plurality of storage devices according to the teachings of the invention.

In a related aspect, the controller further effects transition of a storage device from a power-on mode to a power-off mode if no access request, e.g., no read/write request, is pending for that storage device and a selected time period, e.g., a few seconds, a few minutes, or a few hours, has elapsed since the last read/write request for that storage device.

A variety of storage devices can be utilized in a system according to the invention. Such storage devices include, but are not limited to, magneto disks and optical media. Each storage device can have a data storage density in a range of about 100 Megabytes per cubit centimeter to about 1 Gigabytes per cubic centimeter, and more preferably in a range of about 100 Megabytes per cubic centimeter to about 10 Gigabytes per cubic centimeter. A group, or the entire, of storage devices can be housed in an enclosure (chassis), and a plurality of chassis can be disposed on a rack. The storage devices in a system of the invention can provide, for example, a collective storage in a range of about 25 TB to about 50 TB per chassis and in a range of about 250 TB to about 500 TB per rack. Further, the storage devices can form a RAID storage system. It should be understood that as the storage capacity of storage media suitable for use in a system of the invention increase, the collective storage capacity, or in other words, data storage density, provided by a system of the invention can also increase.

In another aspect, a storage system of the invention as described above, can include a relay coupled to the controller that receives signals from the controller, and electrically connects or disconnects one or more selected ones of the storage devices to a source of power.

In further aspects, a data storage system according to the invention can include a cache storage, having, for example, a cache memory and a cache disk, coupled to the controller for storing selected data retrieved from one or more of the storage devices.

Brief description of the drawings

FIGURE 1 schematically illustrates an exemplary data storage system according to the teachings of the invention,

FIGURE 2 is a block diagram depicting various steps in a method according to the teachings of the invention for managing power distribution to a plurality of storage devices,

FIGURE 3 is a diagram illustrating cost/performance characteristic of an exemplary data storage system of the invention relative to a number of conventional systems,

FIGURE 4 is a diagram schematically depicting an exemplary prototype data storage system built according to the teachings of the invention, and

FIGURE 5 schematically depicts the storage devices of FIGURE 4 housed in an enclosure.

Detailed Description

The present invention provides systems and methods for cost-effective storage and retrieval of a large amount of data while minimizing physical space required for such storage. As discussed in more detail below, a system of the invention can include a plurality of selected storage media, e.g., disks, which can be, for example, packed in an enclosure in close proximity of one another. Each storage medium is normally in a power-off state in order to alleviate the thermal load of the system. A controller is utilized to transition a selected one of the storage media from a power-off mode into a power-on mode in order to read data from and/or write data to that storage medium.

With reference to FIGURE 1, an exemplary data storage system 10 according to the teachings of the invention includes a plurality of storage devices 12, for example, disks, provided in an enclosure 14, and a controller 16 that can communicate with the storage devices 12 via, for example, a bus 18. The controller 16 can be housed within the enclosure 14, or alternatively, it can be provided external to the enclosure. The

storage devices, as described above, that reduces the overall power consumption of the system. This allows a more compact configuration for the storage system, and also allows more disk drives to share the same electronics control system, thereby lowering the cost of manufacturing.

Further, an initial access latency to a storage device that is in a power-off mode in a data storage system of the invention can be approximately 10 seconds. This access latency is comparable to the best case, i.e., tape drive is empty and data is located at the beginning of the tape, access time for robotic tape libraries. Any additional access for performing read/write operations in data storage system of the invention will be at full random access speed.

Moreover, as discussed above, in a system of the invention, the storage devices, e.g., disks, are normally in a power-off state. This advantageously reduces wear and tear experienced by each storage device if it is accessed infrequently, thereby lengthening its shelf life. For example, magnetic disks cease to spin when transitioned into a power-off state, and hence experience less wear and tear in this state.

In some embodiments of the invention, techniques can be utilized to maintain the most frequently accessed drives highly available, for example, by lengthening the inactive period after which the device is transitioned to a power-off mode.

A direct disk peripheral interface can enhance database performance by eliminating the software overhead associated with distributed networked storage. The expected data storage I/O rate can be supported using a high speed interface.

FIGURE 3 schematically depicts the cost/performance characteristics of an exemplary data storage system of the invention having an array of disks relative to those of a number of conventional storage systems. The graph of FIGURE 3 plots performance versus cost (in a log-log scale). As shown in this figure, a data storage system of the invention can provide considerably enhanced performance relative to tape libraries or NAS devices at comparable or reduced cost. Further, a data storage system of invention can be less costly than a conventional RAID system.

In order to demonstrate the feasibility of manufacturing a storage system according to the teachings of the invention, and the efficacy of such a system for storage and retrieval of a large amount of data, a prototype system was built and tested. FIGURE 4 schematically illustrates that this prototype system includes a controller 16

example, more disks can be in a power-on state if they are sparsely distributed. In this exemplary prototype, it is feasible to have about 25 percent of the disks in a power-on state without encountering any thermal overload. It should, however, be understood that this exemplary prototype is provided only as an example, and the 25 percent limit is not intended to indicate an absolute upper limit in other embodiments of the invention. In particular, various improvements, including providing better thermal insulation and/or cooling mechanisms, can be employed to increase the maximum number of disks that can be simultaneously in a power-on state.

When the controller 16 receives a request for access to a disk that is in a power-off state while the number of other disks that are in the power-on state has reached an upper threshold imposed by the thermal load, the controller 16 can suspend access to one of the disks that is already in a power-on state, and transition that disk to a power-off state, in order to allow switching on the requested disk that is in a power-off state. The selection of a disk to be transitioned into a power-off state to allow transitioning a new disk from a power-off state to a power-on state can be performed based on a FIFO protocol, although other protocols can also be employed. In a FIFO protocol, a disk that has been in a power-on state for the longest time period is the first to be selected for being transitioned into a power-off state. If the selected disk is presently processing an input-output (I/O) request, the I/O processing can be blocked before transitioning the disk into a power-off state. The blocked I/O processing can, however, be scheduled to resume once the disk can be switched back on without causing thermal overload, for example, once one or more other disks have been switched off. A scheduler can manage the blocking and resumption of the I/O requests based on a selected scheduling protocol. Such a scheduler can be built, for example, as a kernel process or alternatively as a multi-threaded user program.

With continued reference to FIGURE 4, the exemplary controller 16 is also in communication with a memory cache 40, which can in turn communicate with a disk cache 42 for storing selected data retrieved from any of the hard disks 32. The data stored on the memory cache or the disk cache can be subsequently retrieved, if desired, very rapidly. In this exemplary protocol, when the controller receives a request for retrieval of a portion of a file residing on one of the disks, the controller retrieves the entire file, or an entire directory in which the file resides. The requested portion is

What is claimed is:

1. A data storage system, comprising:

a plurality of storage devices for storing data, and

a controller coupled to the storage devices to effect transition of one or more of the storage devices from a power-off mode to a power-on mode upon receipt of a read/write request for those storage devices, the controller further effecting transition of a storage device from a power-on mode to a power-off mode if no read/write request is pending for that storage device and a selected time period has elapsed since the last read/write request for that storage device.

2. The data storage system of claim 1, wherein the plurality of the storage devices form a RAID storage system.

3. The data storage system of claim 1, wherein each of the plurality of the storage devices is normally in a power-off mode.

4. The data storage system of claim 1, wherein said plurality of storage devices provide a collective data storage capacity in a range of about one hundred Terabytes to a few hundred Terabytes.

5. The data storage system of claim 1, wherein said plurality of storage device provide a collective data storage capacity in a range of tens of Terabytes to a few hundred Terabytes.

6. The data storage system of claim 1, wherein said plurality of storage device provide a collective data storage capacity in a range of about 50 Terabytes to about 100 Terabytes.

7. The data storage system of claim 1, wherein each of said plurality of storage devices provides a data storage density in a range of about 100 Mbytes per cubic centimeter to about 10 Gigabytes per cubic centimeter.

16. The storage system of claim 14, wherein each of said racks provides a storage capacity in a range of about 250 to about 500 Terabytes.

17. The storage system of claim 10, wherein said storage devices comprise any of a magnetic hard disk or an optical storage medium.

18. The storage system of claim 10, further comprising a relay electrically coupled to said controller for receiving signals from said controller to connect or disconnect one or more selected ones of said storage devices to a source of power.

19. The storage system of claim 10, further comprising a cache storage medium in communication with said controller for storing selected data retrieved from one or more of said storage devices.

20. A method for managing power distribution to a plurality of storage devices, effecting transition of a storage device from a power-off mode to a power-on mode upon receipt of a request for writing data to or reading data from that storage device, and effecting transition of a storage device from a power-on mode to a power-off mode if no read/write request is pending for that storage device and a selected time period has elapsed since the receipt of the last read/write request.

21. The method of claim 20, wherein said time period is selected to be in a range of about a few seconds to about a few hours.

22. The method of claim 21, wherein said time period is selected to be in a range of about a few minutes to about a few hours.

FIG. 1

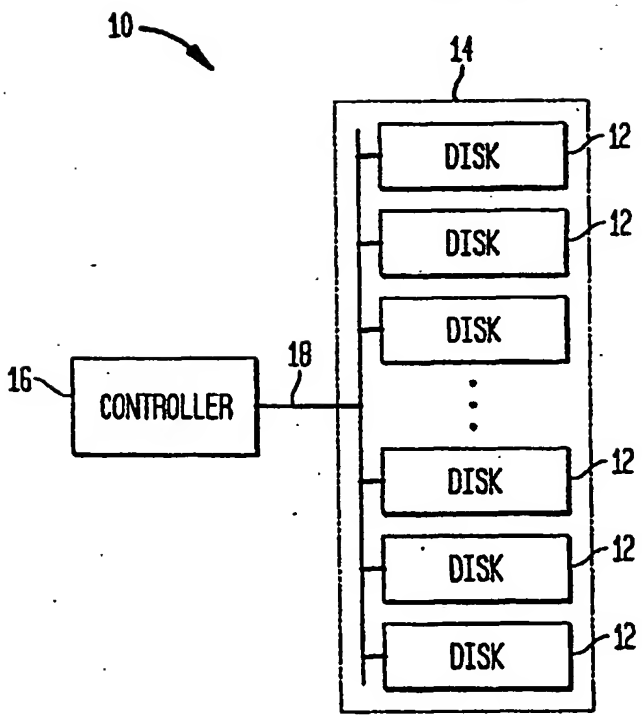

10
14
DISK
12
DISK
12
DISK
18
16
CONTROLLER
DISK
12
DISK
12
DISK
12

FIG. 3

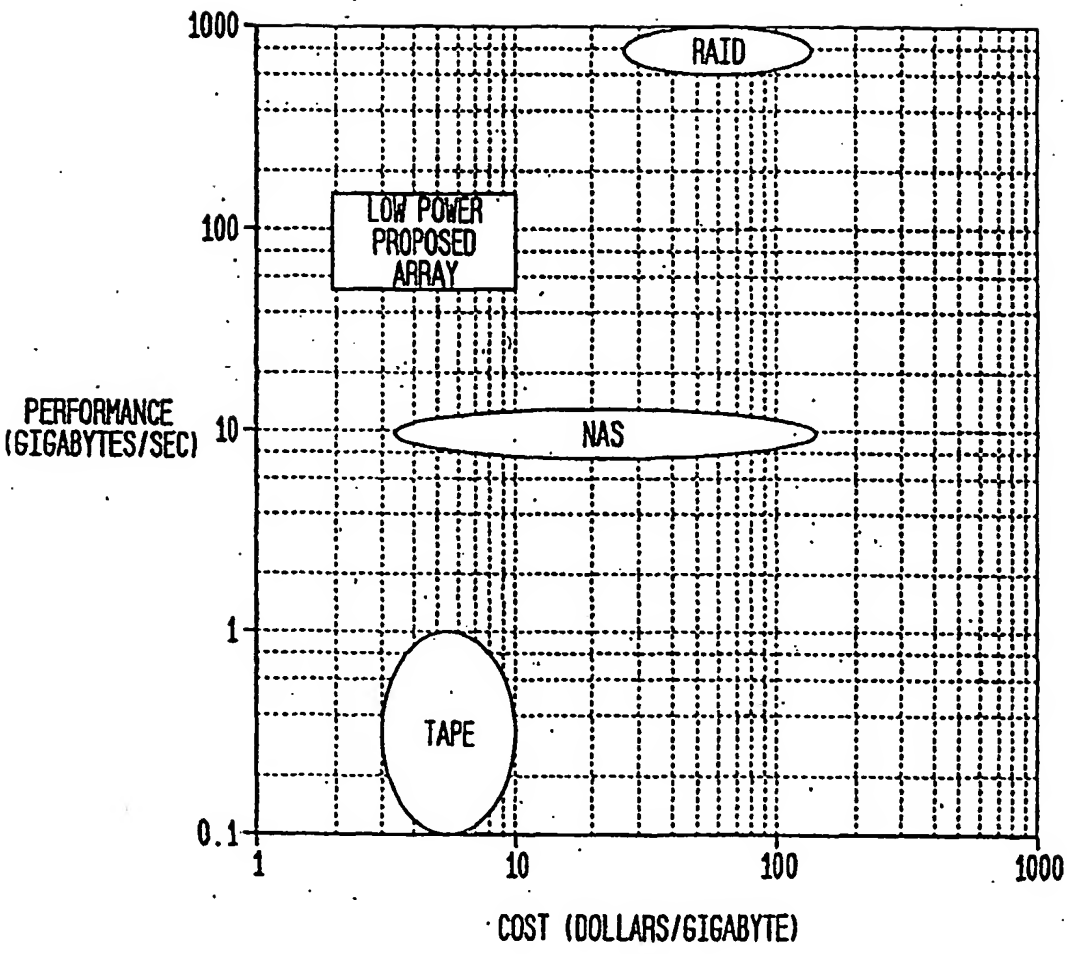

1000
100
10
1
0.1
PERFORMANCE (GIGABYTES/SEC)
RAID
LOW POWER PROPOSED ARRAY
NAS
TAPE
1
10
100
1000
COST (DOLLARS/GIGABYTE)

FIG. 5